Met Office

32-bit Physics in the Unified Model

A proof of concept study

Technical Note No. 626

January 2018

Richard Gilham

Contents

32-bit Physics in the Unified Model....1

Introduction....2

Context and motivation....3

Proof of concept project: Large Scale Precipitation scheme....6

Further work....13

Acknowledgements....13

References....14

© Crown copyright 2018 Met Office

Introduction

The desire to improve the skill of the Met Office's weather forecasts and climate projection is inexorably linked to our capacity to compute the evolution of the Earth System. Advances can be made through improvements to the scientific content of our simulations, or improving our capacity to run these simulations. This work investigates the use of lower precision data in the Unified Model (UM) with a view to improving computational performance.

At present the UM generally uses 64-bit numbers to represent the atmosphere. However, existing software and hardware also supports lower precision 32-bit numbers, which is an attractive proposition for several reasons:

- It provides a fresh opportunity to consider what *information* model fields actually carry and what *data* is required to represent it.
- 32-bit data is generally processed faster than 64-bit data by existing hardware, increasing throughput on present HPC systems.
- It affords more flexibility in the range of hardware that might be leveraged in the future (e.g. GPUs, FPGA), better positioning the UM to take advantage of industry innovation currently being driven by Machine Learning.

This paper documents a proof-of-concept project to determine the scientific and technical feasibility of running parts of the UM that describe sub-grid-scale physical processes (generally referred to as 'physics') at 32-bit precision.

© Crown copyright 2018 Met Office

Context and motivation

The crux of this topic is understanding the difference between information and data, and ensuring that the decisions we make provide a sustainable environment for developing and using numerical weather prediction in the UM. This has implications for all applications of the UM, spanning both weather forecasting and climate simulations.

At present, the UM takes a precautionary approach to choosing the precision of its variables, with 64-bit (double) precision being the default. This corresponds to approximately 17 significant figures[1]. This is a legacy of the vector supercomputers that the UM was developed on, where there was minimal performance penalty for choosing double over single precision. Over the decades and many generations of supercomputers, explicit consideration of data precision has been overlooked and not considered important.

The changing nature of the world's most powerful supercomputers now provides a timely opportunity to re-evaluate this issue. Four general questions set a framework for this:

- What *information* do we need in the model?
- How much *data* do we need to represent this information?
- How data-efficient is the UM now?
- How might we improve data efficiency in the UM?

We have up-to-date knowledge of the information/data issue in two areas. The first relates to how we manage model diagnostics using the so-called 'WGDOS' packing system (UMDP-F03, UMDP-C04). WGDOS packing is a lossy compression algorithm, and is used to great effect to reduce data volumes to find a compromise between scientific need and storage costs. Standard packing profiles are defined for each output field for both weather and climate applications, allowing the appropriate precision to be retained.

[1] IEEE754 provides full information regarding floating-point arithmetic, including the subtleties of representation in binary and decimal formats.

© Crown copyright 2018 Met Office

Table 1 Examples of packed field absolute precision and bits required for common model fields. Indicative ranges for demonstration purposes only.

Model Field	Indicative Range	Packed Accuracy/ Bits Required	
		Operational	Climate
Wind u-component (ms^{-1})	-100 to 100	0.125 / 12 bit	$2.4e^{-4}$ / 20 bit
Specific Humidity ($kgkg^{-1}$)	10^{-8} to 10^{-1}	$6.0e^{-8}$ / 22 bit	Not Packed
Surface Temperature (K)	150 to 350	0.125 / 12 bit	$2.4e^{-4}$ / 20 bit
Large Scale Rain Rate ($kgm^{-2}s^{-1}$)	0 to 0.5	$3.81e^{-6}$ / 18 bit	Not Packed

Table 1 illustrates diagnostic packing for example model fields. The chosen packing level implies the level of precision that provides acceptable accuracy to capture the desired information for the chosen application. Climate generally demands a higher precision to facilitate analysis of subtle long-term trends. An interesting feature of this packing strategy is that the amount of data required to store a field to a chosen relative accuracy depends on it absolute value. This means that the data to information ratio in the model is very dependent on the units used.

In some other fields of computational science, reduced units are used to rescale state variables to magnitudes close to unity, facilitating better data efficiency and numerical stability. Applying these thought processes to state variables actually stored in the UM at runtime may provide a strategic approach to better data efficiency.

The second area we have good knowledge of information/data in the UM is the successful use of 32-bit compute in the UM solver. In this case, the desired convergence criterion was obtainable using 32-bit rather than 64-bit numbers. Although slightly more iterations were required to reach convergence, the faster iteration time resulted in an overall reduction in compute time due to reduced inter-processor communication and better utilization of fast on-chip cache. Table 2 gives an example of the reduction in compute time when using 32 rather than 64-bit numbers.

© Crown copyright 2018 Met Office

Table 2: Computational saving through use of the single precision solver.

	Full Compute (seconds)	Solver (seconds)	Stabilised Bi-conjugate Gradient Algorithm (seconds)
Double (64-bit) Precision	1955	500	370
Single (32-bit) Precision Helmholtz Matrix	1812	355	227
Single (32-bit) Precision	1763	309	181

We may also look to other NWP models for a guide as to the value and interest in carefully managing data precision. ECMWF's IFS model was recently upgraded to use 32-bit numbers, yielding a ~40% speed-up with little scientific impact (Váňa et al (2017)). Experiments using an idealised Lorenz model have demonstrated acceptable climatology using just 6-bit data (Düben et al (2014)). WRF (vn 3.7 onwards) is now available in double and single precision (WRF (2015)). FV3 has been tested on the EDISON supercomputer showing ~40% performance improvement at 32-bit (FV3 (2017)). COSMO, in recent kilometre-resolution global simulations, uses a partial 32-bit implementation (Fuhrer et al (2017)).

© Crown copyright 2018 Met Office

Proof of concept project: Large Scale Precipitation scheme

The various strands of evidence outlined in the previous section provide a strong motivation for exploring these issues within the UM physics codebase. To achieve this, we have completed a proof-of-concept project focusing on the Large Scale Precipitation (LSP) scheme.

This scheme was chosen two main reasons:

- LSP is one of the most expensive physics schemes, both at global and regional (convection-permitting) scales
- The codebase in this area makes for a technically tractable project due to it being reasonably well encapsulated and not too large

The criteria for success for this project were a balance of minimising scientific impact and maximising technical quality. These constraints are especially challenging in an operational model but are essential for maintaining the high quality of both the scientific and technical attributes of the UM.

The first phase of the project focused on developing a suitable approach to making the scheme precision-aware. This was achieved by defining a conceptual ‘bubble’ within which the data precision is controlled. Within this bubble, the precision of all floating-point numbers is controlled. At the bubble interface, all data passing in and out is appropriately managed, and widely used low-level routines were suitably adapted also. Through this design, the desired technical capability was realised whist preserving the readability and maintainability of the science. This is an important aspect of code quality because it helps minimise any technical barriers to future science development.

© Crown copyright 2018 Met Office

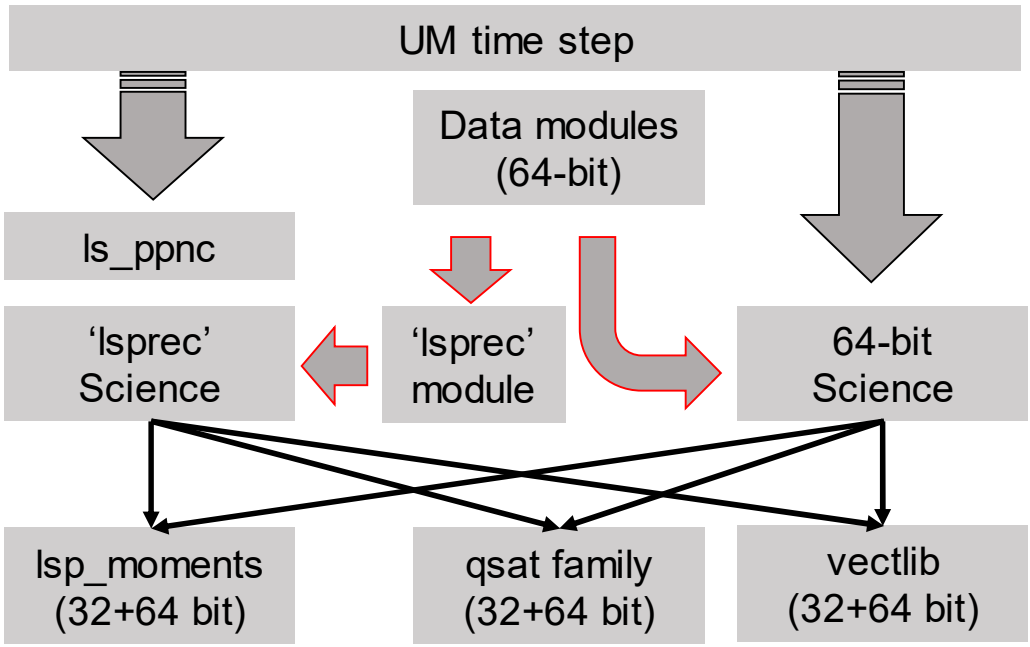

Figure 1: Schematic of the precision-aware Large Scale Precipitation scheme.

The process of implementing this design highlighted a number of technical impediments. These related to historically acceptable programming motifs that were incompatible with the more advanced language features required. Two in particular required attention:

First point passing- passing just the first element of each field and the number of points. This relies on the layout of data in memory. For example:

```
.
.
.
  CALL my_sub(arr1(1),arr2(1),npnts)
.
.
.
SUBROUTINE my_sub(arr1,arr2,npnts)
  IMPLICIT NONE
  INTEGER, INTENT(IN)  :: npnts
  REAL,    INTENT(IN)  :: arr1(npnts)
  REAL,    INTENT(OUT) :: arr2(npnts)
.
.
.
```


© Crown copyright 2018 Met Office

Implicit reshaping- changing the dimensionality of fields through a subroutine interface. This again takes advantage of memory layout, and is sometimes combined with first point passing:

```
.
.
.
  INTEGER, PARAMETER :: npntsi = 128, npntsj=128
  REAL :: arr1(npntsi,npntsj)
  REAL :: arr2(npntsi,npntsj)
.
.
.
  CALL my_sub(arr1,arr2,npntsi*npntsj)
.
.
.
SUBROUTINE my_sub(arr1,arr2,npnts)
  IMPLICIT NONE
  INTEGER, INTENT(IN)  :: npnts
  REAL,    INTENT(IN)  :: arr1(npnts)
  REAL,    INTENT(OUT) :: arr2(npnts)
.
.
.
```

The second phase of the project focused on non-scientific testing of the precision-aware scheme. Leveraging the powerful ‘rose-stem’ test harness, a wide range of model configurations were quickly and rigorously tested. This identified further technical challenges, for which solutions were identified.

Of particular note was a bug in the specific version of the Cray compiler used as default resulting in processor decomposition test failures. This was caused by last-digit variation in the results from POWER (ie X^Y) calculations when the compiler chose to vectorise that region of code.

© Crown copyright 2018 Met Office

The final phase of the project was technical and scientific evaluation. Experience from the other modelling systems described above suggest that a 40% speedup might be achievable.

Table 3 summarises preliminary computational performance results across a range of model runs. All tests show an improvement in runtime. The high compiler optimisation runs are the most representative of operational use and show the best improvements, with around a 5% improvement in total model runtime. The LSP scheme itself shows close to the hoped-for 40% improvement, except for the GA7 runs. GA7, which is not yet operational for weather applications, features a significant change to the science settings for LSP, altering the flow through the LSP code. Under 'safe' optimisation, the savings are more modest. Results will change as further science optimisation developments are implemented in the future.

The variability between results can be attributed to several factors, including the chosen processor decomposition and the meteorology present in these specific test cases. This is because the LSP scheme only operates on grid-boxes under certain meteorological conditions, which in turn alters the computational effort required. Therefore, the fine detail of 'which processors it is raining on' affects the load balance of the whole model in a non-trivial way. The savings in the LSP scheme through reducing data precision are greater than this variability.

Table 3: Preliminary performance benefit measurements of the 32-bit Large Scale Precipitation Scheme on the Met Office Cray XC40 supercomputer.

	N512 GA6 2 day	N48 GA7 AMIP 30 day		UKV 1.5 km 24 hour		
Build/ decomposition	High 10x23	Safe 12x9	Safe 16x9	Safe 8x33	Safe 12x22	High 6x41
UM speed-up	4.5%	1.3%	2.8%	2.5%	3.9%	5.5%
LSP scheme speed-up	33.5%	14.9%	16.3%	38.0%	35.1%	51.4%

© Crown copyright 2018 Met Office

Initial scientific testing using the global N320 case study suite demonstrates minimal scientific impact, suggesting that the scheme does not need 64-bit data to hold its information. The aim here is not to provide a detailed scientific assessment, but rather give a first-look indication of the impact on model evolution.

Figures 2 & 3 are representative plots from an extensive analysis of the global case studies, a standard set of tests for evaluating new science changes. Many metrics are compared against both observations and model analysis for a range of seasons. The differences between the 64 and 32-bit LSP schemes are plotted in the right-hand column. In general, the differences are very small.

When inspecting the model performance in more detail, there is a slight indication of a bias when switching to the 32-bit scheme on a small proportion of the metrics plotted. An example is given in Figure 3. This can mostly be explained by the need to use a new version of the saturated water vapour pressure calculation (qsat). This incorporates a fix for a subtle science bug (and several technical issues) that introduced a small offset in the output of the routines, another example where attention to detail is critical. However, removing this (not shown) still leaves a very small bias that should be noted for further investigation.

In conclusion, the proof-of-concept project has successfully demonstrated that 32-bit physics in the UM is a technically and scientifically feasible proposition. Detailed working knowledge of hidden technical debts has been gained, which will reduce the barriers to further implementation of precision-awareness around the UM.

The new functionality is available in UM vn10.9 (Autumn 2017) onwards with a single click in the user interface.

© Crown copyright 2018 Met Office

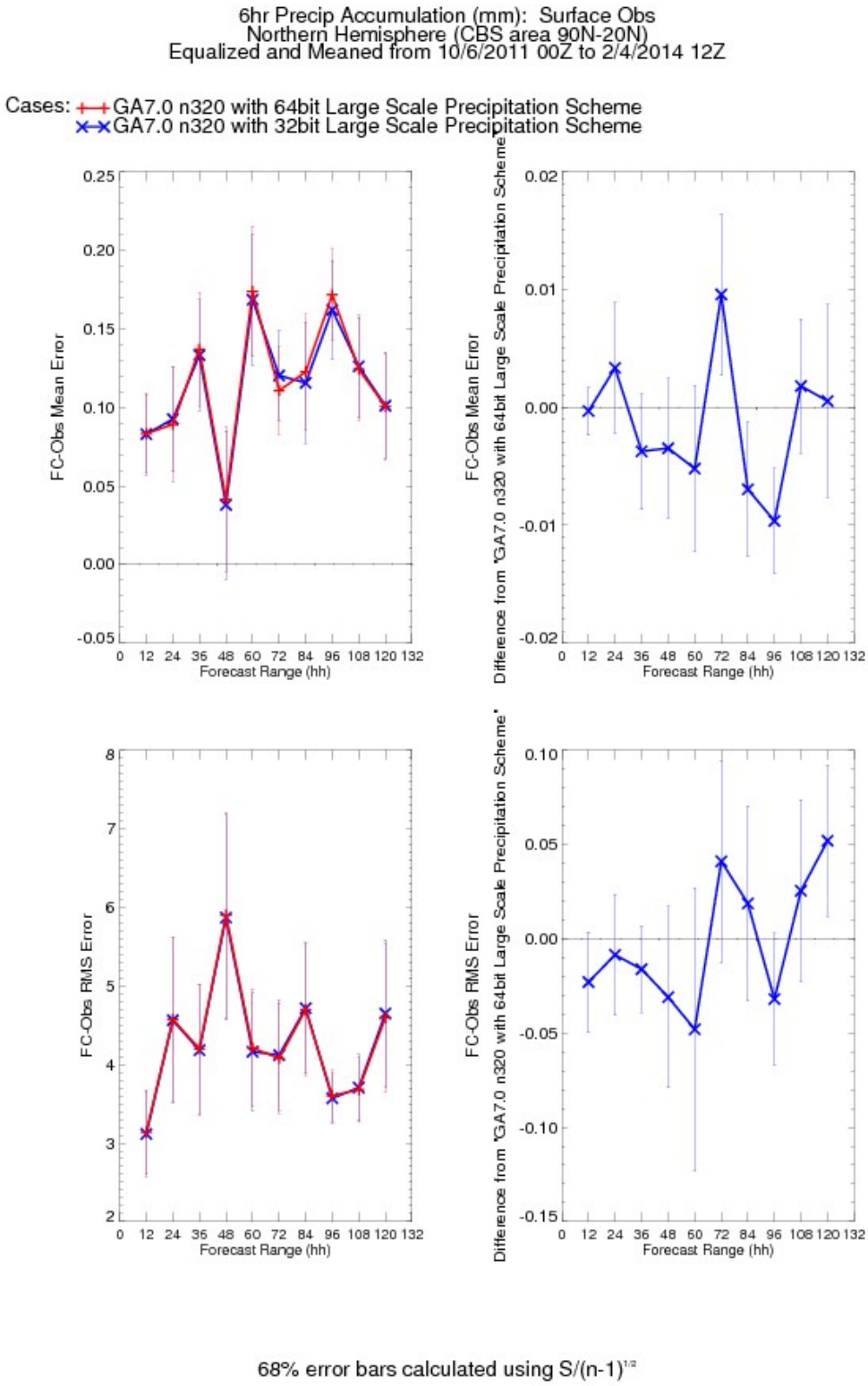

Figure 2: Accumulated Precipitation analysis comparing 64-bit and 32-bit versions of the Large Scale precipitation scheme for a range of N320 case studies. Generated using the TRUI analysis suite.

© Crown copyright 2018 Met Office

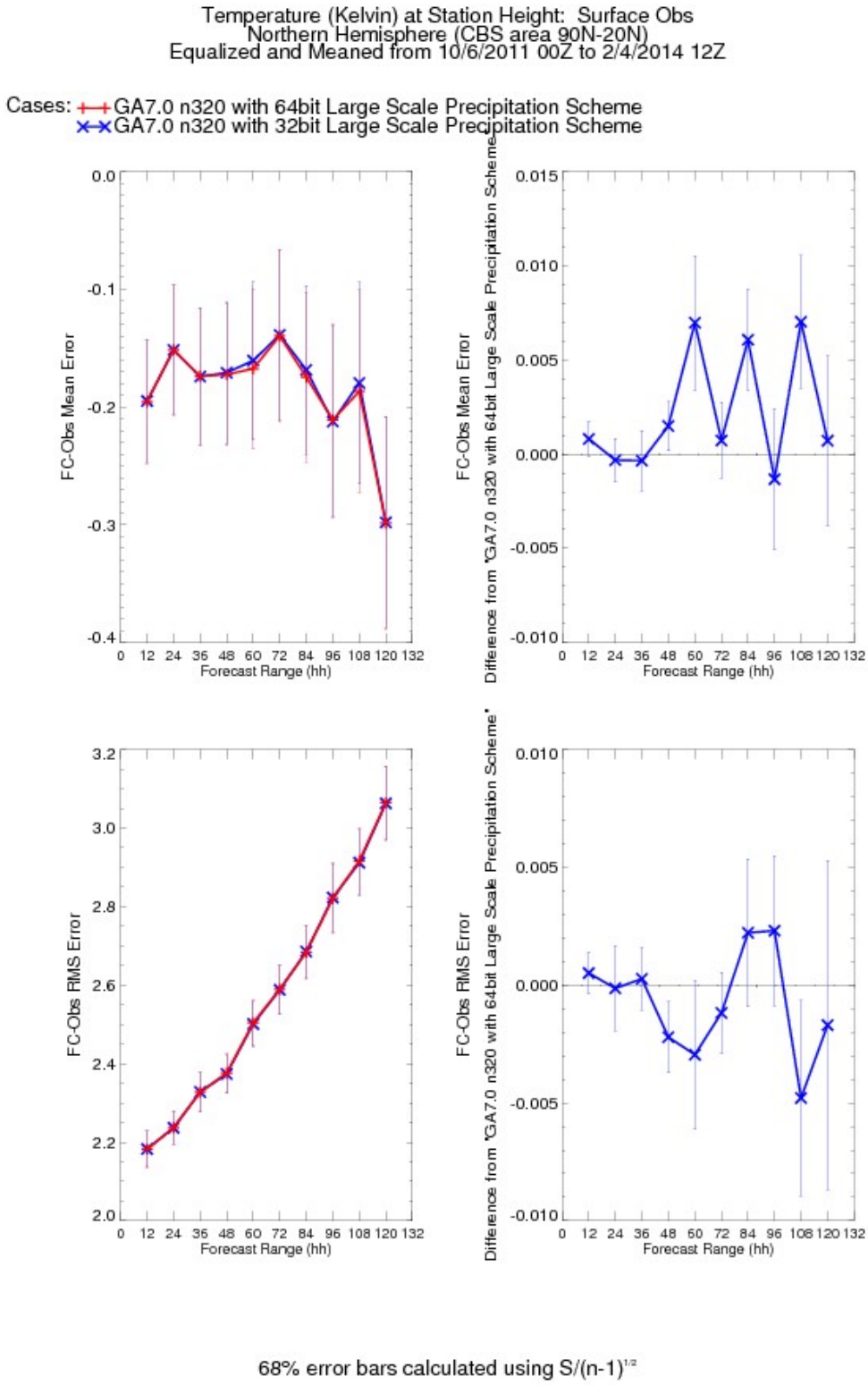

Figure 3: Screen-level temperature analysis comparing 64-bit and 32-bit versions of the Large Scale precipitation scheme for a range of N320 case studies. Generated using the TRUI analysis suite.

© Crown copyright 2018 Met Office

Further work

There are several opportunities for building on the outcomes of this project.

- Further evaluation of the Large Scale Precipitation scheme at 32-bit for inclusion in future model configurations via the Global Atmosphere and Regional Atmosphere projects.
- Application of the same methodology to further physics schemes
- Expansion of the 32-bit 'bubble' to incorporate intermediate control routines, further reducing resource requirements.
- Application beyond the physics schemes, ultimately leading to a holistic approach to precision in the UM.
- Consideration of more novel and disruptive approaches to maximise data efficiency, such as the use of reduced units
- Transfer the outcomes of the project to LFRic, the successor to the UM, helping to ensure 'precision awareness' is engineered into the model at a top level at an early stage.

Acknowledgements

The nature of this work straddles and challenges the traditional classifications of 'science' and 'technical' work, and has therefore drawn on the expertise of both the Scientific Software Engineer and Scientist professions in the office. Specifically, the author would like to thank the HPC optimisation and UM systems teams for extensive support regarding the technical aspects of this work, and Jonathan Wilkinson, the LSP code owner, for supporting a potentially disruptive project on his science.

© Crown copyright 2018 Met Office

References

Düben et al (2014) "*The use of imprecise processing to improve accuracy in weather & climate prediction*" Journal of Computational Physics, Vol. 271 (August 2014), pp. 2-18, doi:10.1016/j.jcp.2013.10.042

Fuhrer et al (2017) "*Near-global climate simulation at 1 km resolution: establishing a performance baseline on 4'888 GPUs with COSMO 5.0*" Geosci. Model Dev. Discuss. IN REVIEW doi: 10.5194/gmd-2017-230

FV3 (2017): www.gfdl.noaa.gov/fv3/fv3-performance

UMDP-F03: Unified Model Documentation Paper F03 Input and Output File Formats[2]

UMDP-C04: Unified Model Documentation Paper C04 Storage Handling and Diagnostic System (STASH)

Váňa et al (2017) "*Single Precision in Weather Forecasting Models: An Evaluation with the IFS*" Mon. Wea. Rev., Vol. 145, No. 2. (7 December 2016), pp. 495-502, doi:10.1175/mwr-d-16-0228.1

WRF (2015): http://www2.mmm.ucar.edu/wrf/users/wrfv3.7/updates-3.7.html

[2] UM Documentation Papers are readily available to licenced UM users, and also to others on request.

© Crown copyright 2018 Met Office

Met Office
FitzRoy Road, Exeter
Devon EX1 3PB
United Kingdom

Tel (UK): 0370 900 0100 (Int): +44 1392 885680
Fax (UK): 0370 900 5050 (Int): +44 1392 885681
enquiries@metoffice.gov.uk
www.metoffice.gov.uk